DOCUMENT RESUME

ED 421 536 TM 028 862

AUTHOR Kromrey, Jeffrey D.; Parshall, Cynthia G.; Yi, Qing
TITLE The Effects of Content Representativeness and Differential Weighting on Test Equating: A Monte Carlo Study.
PUB DATE 1998-04-00
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Equated Scores; High Schools; *Item Response Theory; Monte Carlo Methods; *Test Content; Test Items
IDENTIFIERS ACT Assessment; Anchor Tests; *Weighting (Statistical)

ABSTRACT

The effects of anchor test characteristics in the accuracy and precision of test equating in the "common items, nonequivalent groups design" were studied. The study also considered the effects of nonparallel based and new forms on the equating solution, and it investigated the effects of differential weighting on the success of equating under these conditions of nonrepresentative anchor tests and nonparallel test forms. Data were generated for this simulation study using a multidimensional item response theory approach to data from the American College Testing Program assessment in mathematics. The three weighting models included the traditional unweighted approach and two differential weighting methods, one with item weights obtained using the proportion of items in the anchor test relative to the proportion in the total test and one based on the proportion of item response theory information provided in each content area in the anchor test relative to the proportion in the total test. For the conditions examined in this study, the traditional unit weighting method outperformed both alternative methods. Despite the limited performance of the alternative methods in this study, they merit further study. (Contains 8 tables, 4 figures, and 17 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 421 536

The Effects of Content Representativeness and Differential Weighting on Test Equating: A Monte Carlo Study

Jeffrey D. Kromrey
Cynthia G. Parshall
Qing Yi

University of South Florida

Cynthia Parshall

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028862

Paper presented at the annual meeting of the American Educational Research Association, April 13-17, 1998, San Diego, CA.

The Effects of Content Representativeness and Differential Weighting on Test Equating: A Monte Carlo Study

Test equating is a set of procedures designed to remove the effects of test form differences from sets of test scores so that the scores obtained from different forms may be placed onto the same score scale. A variety of equating designs have been described by Angoff (1984) and Peterson, Kolen and Hoover (1989). A frequently used design in many testing programs is the common-item, non-equivalent groups design (often referred to as the Angoff Model IV). In this design, one form of an examination is given to one group of examinees, and a second form is given to a second group of examinees. In addition, a common form (called an anchor test) is given to both groups of examinees. In this design, no assumption is made about the equivalence of the groups on the attribute being measured.

This equating design has certain advantages over other, commonly used designs. For example, in the single group design, all examinees are administered two exam forms. In the random group design both forms are also administered, typically by giving alternate examinees at a test site one of the two forms. However, in the common-item, non-equivalent groups design, only a single test form needs to be administered at a given test date and each examinee takes only a single exam. Any differences in group ability or test form difficulty are identified and controlled through the use of the anchor set of items.

The anchor test is clearly of critical importance in obtaining an accurate equating relationship. The relative effectiveness of anchor tests on equating solutions has been investigated under a variety of conditions (Cook & Peterson, 1987; Hills, Subhiyah, & Hirsch, 1982; Holmes, 1982; Motika & Chason, 1995; Norcini, Shea, & Lipner, 1994). Petersen, Marco, and Stewart (1982) investigated a variety of test form and anchor test characteristics, including the similarity of the anchor test to the total tests in terms of content and item difficulty. They consistently found both of these anchor test properties to affect the quality of equating when the examinee groups differed in ability. The length of the anchor test has also been related to the success of the equating relationship (Budescu, 1985), particularly when examinee groups differ in ability (Klein & Kolen, 1985). However, a more critical anchor test characteristic may be the extent to which the anchor reflects the content of the total test. Klein and Jarjoura (1985) found a content representative anchor to function better than a longer, non-representative anchor. Budescu (1985) suggested that the most efficient equating results are derived when the correlation between anchor test and total test is high, and when a test is constructed such that anchor items comprise one half of its total length. Cook and Petersen (1987) summarized several studies that considered anchor test properties, including anchor test length, the parallelism of the anchor test to the total test forms, and the consistency of item difficulty across forms. Their summary was that the effectiveness of an anchor test is dependent on the extent to which the anchor test is similar to the total test.

Anchor test representation may be compromised if items on the anchor must be discarded because of changes to the content domain that alters the correct response or because of technical difficulties with the items. An additional problem arises when, over the course of time, test

forms change. In many applied educational and licensure test programs, changes in the knowledge base being assessed necessitate changes in test content to maintain the validity of the assessment. Such changes lead to conditions in which the content balance of the new test form is no longer identical to the content balance of the base form (Brennan & Kolen, 1987). And, the anchor tests based on old test forms are no longer strictly representative of the new test forms.

A final area in which issues of content representativeness may arise is that of customized norm-referenced testing (Allen, Ansley & Forsyth, 1987; Way, Forsyth & Ansley, 1989; Linn & Hambleton, 1991). In customized norm-referenced testing, school districts may substitute locally constructed test items for selected items on a nationally normed test (NRT), may drop sections of the NRT, or may augment the NRT with locally developed curriculum-specific tests. Despite these changes to the test content, valid normative information may still be sought, information that requires linkage to the national sample of the NRT.

The accuracy of test equating conducted under all of these conditions is suspect, but little empirical evidence is available to guide practitioners in their judgments about the viability of test equating under these conditions of nonrepresentative anchors.

Most of the previous research on anchor tests and content representativeness has used equal item weights to derive the anchor scores used in the equating. With representative anchors, the use of equal weights is appropriate. However, under conditions of nonrepresentative anchors, differential weighting of anchor items may yield better equating solutions than that obtained with equal item weights. That is, the relative weights applied to anchor items may be used to partially adjust for differences in content between the anchor test and the total test form. Harris (1991) considered the situation in which an anchor test is not fully representative of the content of the total test forms, and then investigated the effects of weighting nonrepresentative anchor items. The anchor test items for each content category were weighted by proportion of items in that category on the total test. Two test links were examined; in one, the weighting procedure performed best, while in the second, the unweighted equating was superior.

Purpose

Many testing programs must make modifications to examinations in order to keep up with changing content areas. In other applications, such as the development of customized norms in achievement tests, problems with a lack of content representation also have been reported. Despite these content changes, in order to maintain testing programs, it is necessary to link assessments. This research informs testing practitioners of the extent to which content changes affect equating accuracy and provides a preliminary examination of the use of differential weighting in test equating.

The purpose of this study was to investigate the effects of anchor test characteristics on the accuracy and precision of test equating in the common items non-equivalent groups design. Further, the study considered the effects of non-parallel base and new forms on the equating solution. Finally, the study was designed to investigate the effects of differential weighting on

the success of equating under these conditions of nonrepresentative anchor tests and non-parallel test forms.

Methods

Data were generated for this simulation study based on a multidimensional item response theory (MIRT) approach. This approach models the complexity and variability inherent in real data by allowing a richer measure of the multiple skills and abilities that examinees often use in responding to an assessment task. The simulation model includes not only the major dimension that provides the basic structure for a given exam, but also includes the numerous minor dimensions that are characteristic of actual data. Thus, MIRT data generation provides simulated data that are more similar to real data than are unidimensionally simulated data (Davey, Nering, & Thompson, 1997; Parshall, Kromrey, Chason, & Yi, 1997).

This method of generating data begins by fitting a multidimensional latent trait model to a large sample of actual data. In this case, a set of 480 items from the ACT Mathematics program were utilized. Approximately 3,500 examinee responses per item were available. The set of items was calibrated using a modified version of NOHARM (Fraser, 1986; Fraser & McDonald, 1988), in order to fit 50 dimensions to the data. No attempt was made to interpret the resulting solution; rather, the fitted model was treated as a template for generating new data.

Items on the ACT Mathematics Test are classified according to three general content categories. These categories are: Pre-Algebra/Elementary Algebra, Intermediate Algebra/Coordinate Geometry, and Plane Geometry Trigonometry. These content areas were used to generate tests assembled according to the study design, and were not reflective of the actual ACT Mathematics test specifications.

Pairs of test forms were constructed from this set of 480 items. For the investigation into anchor representativeness, levels of total test length, anchor test length, and content representativeness were varied. Total test lengths of 30, 60, and 120 items were used. Anchor test lengths of 20%, 33%, and 50% of the total test length were also investigated. Finally, the extent to which these anchors were representative of the content on the total test was also varied. For the conditions in which test forms were parallel but anchors were not necessarily representative, anchor tests were constructed that were proportionally representative of the total test, mildly disproportional, and severely disproportional. In the mildly disproportional, or nonrepresentative case, all three test domains were represented in the anchor test, although two of the domains were under-represented and one was over-represented. In the more severely nonrepresentative condition, one of the domains was missing from the anchor test, one was over-represented in the anchor test, and one was under-represented. The combination of variables investigated resulted in a 3 x 3 x 3 design, for a total of 27 conditions.

For the investigation into the effect of non-parallel test forms on equating success, levels of total test length, anchor test length, and test form parallelism were varied. For this aspect of the study, 27 conditions were also investigated. The same three test lengths and three anchor test lengths described above were used. Tests were also constructed that were parallel across base

and new forms, that were non-parallel to a minor extent, and that were non-parallel to a major extent. Minor non-parallelism was reflected by test forms in which all three domains were present, but one was over-represented on the new form and two were under-represented. In the major non-parallel condition, one domain was not represented on the new form, one was over-represented and one was under-represented. (The content representative, parallel forms case was thus the same for both investigations.)

For each condition examined in the study, 1000 pairs of test forms (e.g., forms A and B) were assembled by randomly selecting items from the archived data files. Once the test forms were assembled, item responses were generated for examinees on forms A and B, using 1000 examinees for each test form. For each pair of test forms, linear equating was conducted, using three weighting methods. The three weighting methods included the traditional unweighted approach (using equal weights for all items in the anchor test), and two differential weighting methods. For one of these, item weights were obtained using the proportion of items in the anchor test relative to the proportion in the total test. For the final weighting method, item weights were based on the proportion of IRT information provided in each content area in the anchor test relative to the proportion in the total test.

Items weights for the proportional weighting method are given by

$$WP_D = \frac{N_{TD}\, N_U}{N_{UD}\, N_T}$$

where

WP_D	is the weight applied to items in domain D,
N_{TD}	is the number of items in domain D on the total test,
N_{UD}	is the number of items in domain D on the anchor test U,
N_U	is the total number of anchor items, and
N_T	is the total number of test items.

Items weights for the information weighting method are given by

$$WI_D = \frac{I_{TD}\, N_U}{I_{UD}\, Q}$$

where

WI_D	is the weight applied to items in domain D,
I_{TD}	is the sum of the information in domain D on the total test,
I_{UD}	is the sum of the information in domain D on the anchor test U,
N_U	is the total number of anchor items, and
Q	is a scaling factor given by

$$Q = \sum \frac{N_{TD} I_{TD}}{I_{UD}}$$

where the summation is over test domains.

The linear equating method was presented by Angoff (1984), in which the conversion of scores on one test form to the score scale of another test form can be accomplished linearly. The equated score on test form A for examinees who have taken form B can be calculated by the following equation

$$Y = AX + B = \frac{S_{bt} X}{S_{at}} + \frac{(M_{bt} - S_{bt} M_{at})}{S_{at}}$$

where

Y represents the equated score on form A for examinees who have taken form B,

X is the test score on form B,

A indexes the slope, $A = \frac{S_{bt}}{S_{at}}$, and

B represents the intercept of the linear equation, $B = \frac{M_{bt} - S_{bt} M_{at}}{S_{at}}$.

The means and standard deviations of the two test forms in the total group can be estimated. For test form A

$$M_{at} = M_{ax} + \frac{r_{aux} S_{at}}{S_{ux}} (M_{ut} - M_{ux}) \text{ and } S_{at}^2 = S_{ax}^2 + \frac{r_{aux}^2 S_{at}^2}{S_{ux}^2} (S_{ut}^2 - S_{ux}^2)$$

where

M_{at} is the mean of test form A for total examinee group T,

M_{ax} is the mean of test form A for examinee group X,

M_{ut} is the mean of anchor test U for total examinee group T,

M_{ux} is the mean of anchor test U for examinee group X,

S_{at} is the standard deviation of test form A for total examinee group T,

S_{ax} is the standard deviation of test form A for examinee group X,

S_{ut} is the standard deviation of anchor test U for total examinee group T,

S_{ux} is the standard deviation of anchor test U for examinee group X, and

r_{aux} is the correlation coefficient between test form A and anchor test U.

For test form B, the mean and standard deviation on the total group can be obtained by the similar equations

$$M_{bt} = M_{by} + \frac{r_{buy} S_{by}}{S_{uy}} (M_{ut} - M_{uy}) \text{ and } S_{bt}^2 = S_{by}^2 + \frac{r_{buy}^2 S_{by}^2}{S_{uy}^2} (S_{ut}^2 - S_{uy}^2)$$

where

M_{bt}	is the mean of test form B for total examinee group T,
M_{by}	is the mean of test form B for examinee group Y,
M_{ut}	is the mean of anchor test U for total examinee group T,
M_{uy}	is the mean of anchor test U for examinee group Y,
S_{bt}	is the standard deviation of test form B for total examinee group T,
S_{by}	is the standard deviation of test form B for examinee group Y,
S_{ut}	is the standard deviation of anchor test U for total examinee group T,
S_{uy}	is the standard deviation of anchor test U for examinee group Y, and
r_{buy}	is the correlation coefficient between test form B and anchor test U.

The combination of equating and weighting methods resulted in three equating solutions for each of the conditions investigated in this study. These solutions were then compared to "truth", which was defined as the expected true score on form A for examinees who were administered form B. Because the item and person parameters in the simulation study were known, for each examinee completing form B, the expected score on form A was also known. This expected true score was computed using a simulated examinee's known ability levels and the form A MIRT item parameters used to generate the data. The expected true score can be obtain by

$$E_j = \sum_i P$$

where P is defined as

$$P(u_i = 1|\theta_j) = c_i + (1 - c_i) \frac{\exp[1.702\, \mathbf{a}_i' \theta_j + d_i]}{1.0 + \exp[1.702\, \mathbf{a}_i' \theta_j + d_i]}$$

where

u_i	is examinee's score (0/1) on item i $(i = 1, 2, 3, \ldots, n)$,
$\mathbf{a}_i$	is the vector of item discrimination parameters $(a_{ik} = a_{i1}, a_{i2}, a_{i3}, \ldots, a_{im})$ for item i in k dimensions $(k = 1,2,3,\ldots,m)$,
d_i	is the scalar difficulty parameter for item i,
c_i	is the scalar lower asymptote parameter for item i,
θ_j	is the vector of $\hat{\theta}$ for person j $(j = 1,2,3,\ldots,N)$, and
$P(u_i = 1\|\theta_j)$	is the probability of an examinee j correctly answering item i.

The relative success of the test equating was determined by examining the root mean squared error, bias and standard error of equating under each condition examined. Root mean squared error (RMSE) represents the square root of the squared difference between the equated

score and the expected score on form A for examinees who have taken form B. RMSE can be expressed as

$$RMSE = \left[\frac{1}{n_s} \sum_k \frac{1}{n_e} \sum_j (Y_j - E_j)^2 \right]^{\frac{1}{2}}$$

where

k	represents the sample of examinees,
j	indexes individual examinees,
n_s	is the number of samples,
n_e	is the number of examinees per sample,
Y_j	is the equated score on form A for examinee j who has taken form B, and
E_j	is the expected score on form A for examinee j who has taken form B.

RMSE can be partialed into two piece of information, that is, the bias in equated scores and the standard error of equating. Both of these were also examined.

The magnitude of the difference between the expected true score on form A and the equated score on form A (based on the examinee's performance on form B and the sample equating) represents equating error. The bias in equating was obtained as the mean of these differences across examinees and across samples. The bias in the equating can be calculated by the following equation

$$Bias = \frac{1}{n_s} \sum_k \frac{1}{n_e} \sum_j (Y_j - E_j)$$

where all the symbols are defined the same as above.

The standard error (SE) of equating can be obtained by

$$SE = \left[RMSE^2 - Bias^2 \right]^{\frac{1}{2}}.$$

Because different test lengths were examined in this study, the estimates of RMSE, bias and SE were divided by the number of items on the test form. Thus, these statistics are expressed as RMSE per item, bias per item, and SE per item.

The simulation was conducted using SAS/IML, version 6.12. The random number were generated using both RANNOR and RANUNI functions, with a different seed used for each execution of the simulation. The accuracy of the program code was verified using benchmark data sets and hand calculations.

Results

Tables 1 to 3 summarize the RMSE, the bias, and the standard error of equating under the conditions in which test forms were parallel, but anchor representativeness varied. The means and standard deviations of sample RMSE are listed in Table 1. The three weighting procedures functioned very similarly under the content representative conditions. The values of RMSE ranged from 0.038 to 0.082. As the test length increased, so did the accuracy of the estimated equating scores for these three methods. The unit weight method performed better than the two alternative weighting methods under both the moderate and severe non-representative conditions. The information weighting method functioned better than the proportional weighting procedure when the content non-representativeness was severe, however, these two alternative weighting methods performed similarly under the moderate non-representative conditions. The RMSE values ranged from 0.045 to 0.099 when the information weighting method was used, and they ranged from 0.058 to 0.124 for the proportional weighting procedure under the severe nonrepresentative conditions.

Similar results were obtained in the analyses of the bias of equated scores. Table 2 reports these results, while Figure 1 provides a plot of the values averaged across test length and anchor percent. Under the content representative conditions, these three weighting procedures performed very similarly. There were very small to negligible amount of bias under the content representative conditions for these three weighting procedures. The bias values were around zero for the unit weighting procedure. For the proportional weighting method, none of the absolute value for the bias was over 0.002 (that is, 0.2 points on a 100 item test). The bias values for the information weighting method ranged from zero to 0.004 (e.g., 0.4 points on a 100 item test). Both positive and negative bias values were observed when content representativeness was present. Under the two non-representative conditions, the unit weighting procedure appeared to perform better than the two alternative weighting methods. For both of these non-representative conditions, there were still very small to negligible amount of bias when unit weighting was used, and the bias values were around zero. The information weighting method performed better than the proportion weighting procedure when the non-representation was severe, but these two alternative weighting methods functioned similarly under the moderate non-representative conditions. Both positive and negative bias values were obtained for the unit weighting method under these two non-representative conditions, but only negative bias was observed for the two alternative weighting methods.

Table 3 lists the standard error of equating, and Figure 2 displays the average standard error across test length and anchor percent. These three weighting procedures performed very similarly when content was representative. The values of the standard error of equating ranged from 0.038 to 0.081. Under the two non-representative conditions, the unit weighting procedure still outperformed the proportional and information weighting methods. Interestingly, information weighting functioned slightly worse than proportional weighting when the content non-representativeness was severe. The values of the standard error of equating ranged from 0.044 to 0.098 for the information weighting procedure, while they ranged from 0.041 to 0.085 for the proportion weighting method under the severe non-representative conditions. These two

alternative weighting methods performed very similarly under the moderate non-representative conditions.

Tables 4 to 6 present the RMSE, the bias, and the standard error of equating obtained under the conditions of non-parallel test forms. First, an examination of the RMSE (Table 4) suggests that the use of unit weights led to smaller RMSEs relative to proportional or information weights under both minor and major non-parallelism of test forms. The increase in RMSE with the use of non-unit weights ranged from trivially small (e.g., minor degree of non-parallelism, 60-item test with 20% anchors, in which the RMSE increased from 0.058 to 0.059 with both methods of non-unit weighting) to substantial (e.g., major degree of non-parallelism, 30-item test with 50% anchors, in which the RMSE increased from 0.079 to 0.102 with proportional weights). For the major degree of non-parallelism, the information weighting was consistently superior to the proportional weighting, but the results were very similar for the minor degree of non-parallelism. However, in none of the conditions was the use of non-unit weights superior to the results observed with unit weighting.

The bias in equating obtained under the conditions of non-parallel test forms is presented in Table 5 and in Figure 3. The results are congruent with those obtained for non-representative anchors. When unit weights were used for the anchor test, the bias in equating was minimal regardless of the degree of non-parallelism of the test forms. However, the use of proportional weights or information weights introduced a degree of bias in the obtained equating equations, bias which increased as the degree of non-parallelism of the forms increased. For the conditions representing a major degree of non-parallelism, the use of proportional weights resulted in bias that was consistently in a negative direction (underestimating the expected score on the base form), while positive biases were observed for the information weights. For the conditions representing a minor degree of non-parallelism, both proportional and information weighting resulted in positive bias

Finally, the results observed for the standard errors of equating (Table 6 and Figure 4) were also similar to those observed with the nonrepresentative anchor conditions. As expected, the standard errors decreased with increasing test length and increased with increasing degree of non-parallelism of the test forms. In most conditions, the use of proportional weights or information weights increased the standard error slightly. For example, with the minor degree of non-parallelism, a test length of 30 items and 20% anchors, the use of proportional weights increased the standard errors from 0.081 to 0.083, while the use of information weights increased the standard error to 0.084. Note, however, that under the major degree of non-parallelism with the same test condition, the use of proportional weights led to a decrease in standard error of equating (from 0.082 with unit weights to 0.081 with proportional weights).

The correlations between the expected true scores and the equated scores were also examined (see Tables 7 and 8). This correlation indicated the relationship between the expected true scores and the equated scores under studied conditions. Table 7 lists the correlations obtained under the parallel test forms conditions. Across the weighting procedures, there was very little difference in the correlation when there was content representativeness present. The values of the correlation ranged from .900 to .976 for the content representative conditions.

However, when there was content non-representativeness in the anchor tests, the correlation between the expected true scores and the equated scores differed across the weighting methods. The correlation was higher for the unit weighting procedure than for the two alternative weighting methods. The values of the correlation in non-representative content conditions for the unit equating procedure were very close to the values in content representative conditions. However, the correlation values for the two alternative weighting methods were lower under non-representative conditions than the correlations obtained in content representative conditions. Thus, the relationship between the expected scores and the equated scores was distorted when proportional or information weighting method was used. For the non-parallel test form conditions (see Table 8), however, the correlations between the expected scores and the equated scores across all the conditions for these three weighting procedures were very similar.

Discussion

For the conditions examined in this study, the traditional unit weighting method outperformed both alternative methods. It is encouraging to note, however, that the unit method is performing quite well, even under conditions of severely nonrepresentative anchor tests and major non-parallelism of test forms.

The two alternative weighting methods perform similarly to the unit weighting method when the anchor tests were representative of the total test content and when the base and new test forms were parallel. This is a reasonable result, given that the weighting is intended to restore this content representation. Under nonrepresentative conditions, the information weighting procedure shows a slight advantage over the proportional method, although neither produces the improvement in equating solutions sought.

Despite the limited performance of the alternative procedures in this study, further study needs to be conducted. The literature suggests that problems with the common-item, non-equivalent groups equating design may be more critical under additional conditions, not examined in this study. When the two examinee groups are dissimilar in overall ability, characteristics of the anchor test have been found to have a greater impact on the success of the equating (Klein & Kolen, 1985; Cook & Petersen,). Some researchers have also suggested that equating error from nonrepresentative anchors can be much greater when the two groups specifically differ on one or more areas of content (Klein & Jarjoura, 1985; Harris, 1991), particularly if that area of content suffers from the nonrepresentativeness. For this study, groups did not differ in ability by design, but were randomly equivalent. More problematic testing conditions could result in much greater equating error than was found here. Given the need for improvements in equating under these more challenging conditions, it is important that these alternative weighting methods be investigated as potential solutions.

References

Allen, N. A., Ansley, T. N. & Forsyth, R. A. (1987). The effect of deleting content-related items on IRT ability parameters. *Educational and Psychological Measurement, 47*, 1141-1152.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprint of chapter in R. L. Thorndike (Ed.), *Educational Measurement* (2nd Ed.). Washington, DC: American Council on Education, 1971).

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11*, 279-290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*, 13-20.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.

Davey, T., Nering, M. L., & Thompson, T. D. (June, 1997). *Realistic simulation procedures for item response data*. In Miller, T (Chair), High-Dimensional Simulation of Item Response Data for CAT Research. Symposium conducted at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Harris, D. J. (1991). Equating with nonrepresentative common item sets and nonequivalent groups. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Hills, J. R. , Subhiyah, T. G. & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. *Journal of Educational Measurement, 25*, 221-231.

Holmes, S. E. (1982, March). *The effects of test content match and number of items on the accuracy of trait estimates from tests equated with the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation. For common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*, 197-206.

Klein, L.W., & Kolen, M. J. (1985). *Effect of number of common items in common-item equating with nonrandom groups*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R. L. & Hambleton, R. K. (1991). Customized tests and customized norms. *Applied Measurement in Education, 4*, 185-207.

Motika, R. T., & Chason, W. M. (1995, April). *Performance of Angoff Model IV linear test equating using total test and content dimensional sub-test designs in small groups of examinees.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Norcini, J., Shea, J., & Lipner, R. (1994). The effect of anchor item characteristics on equivalent cutting scores. *Applied Measurement in Education, 7*, 187-194.

Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd Ed.) (pp. 221-262). New York: Macmillan.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.) *Test Equating*. New York: Academic Press, 71-135.

Way, W. D., Forsyth, R. A. & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without content sampling. *Applied Measurement in Education, 2*, 15-35.

Table 1
Means and Standard Deviations of Sample Equating RMSE Under Parallel Forms and Across Levels of Anchor Representativeness.

Test Length	Anchor Percent		Anchor Representativeness: Representative			Non-Representative (Moderate)			Non-Representative (Severe)		
			Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt
30	20%	Mean RMSE	0.081	0.081	0.082	0.081	0.084	0.085	0.081	0.090	0.087
		SD RMSE	0.006	0.006	0.006	0.006	0.006	0.007	0.006	0.007	0.007
60	20%	Mean RMSE	0.057	0.057	0.057	0.057	0.063	0.064	0.057	0.070	0.061
		SD RMSE	0.003	0.003	0.004	0.003	0.004	0.006	0.003	0.005	0.004
120	20%	Mean RMSE	0.040	0.040	0.040	0.039	0.044	0.044	0.040	0.058	0.045
		SD RMSE	0.002	0.002	0.002	0.002	0.003	0.003	0.002	0.005	0.003
30	33%	Mean RMSE	0.079	0.079	0.080	0.079	0.096	0.099	0.079	0.110	0.099
		SD RMSE	0.005	0.005	0.005	0.005	0.007	0.010	0.005	0.012	0.011
60	33%	Mean RMSE	0.056	0.056	0.056	0.055	0.069	0.069	0.055	0.092	0.070
		SD RMSE	0.003	0.003	0.003	0.003	0.006	0.006	0.003	0.009	0.006
120	33%	Mean RMSE	0.039	0.039	0.039	0.039	0.050	0.050	0.039	0.082	0.050
		SD RMSE	0.002	0.002	0.002	0.002	0.004	0.004	0.002	0.007	0.004
30	50%	Mean RMSE	0.077	0.077	0.077	0.077	0.106	0.110	0.077	0.124	0.087
		SD RMSE	0.005	0.005	0.005	0.005	0.010	0.015	0.005	0.010	0.006
60	50%	Mean RMSE	0.054	0.054	0.054	0.054	0.075	0.075	0.054	0.113	0.062
		SD RMSE	0.003	0.003	0.003	0.003	0.008	0.007	0.003	0.007	0.004
120	50%	Mean RMSE	0.038	0.038	0.038	0.038	0.057	0.057	0.038	0.106	0.045
		SD RMSE	0.002	0.002	0.002	0.002	0.007	0.006	0.002	0.005	0.003

Table 2
Means and Standard Deviations of Sample Bias in Equated Scores Under Parallel Forms and Across Levels of Anchor Representativeness.

			Anchor Representativeness								
			Representative			Non-Representative (Moderate)			Non-Representative (Severe)		
Test Length	Anchor Percent		Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt
30	20%	Bias	0.000	0.002	0.004	0.000	-0.008	-0.006	-0.000	-0.040	-0.006
		SE	0.006	0.006	0.007	0.005	0.010	0.012	0.005	0.011	0.012
60	20%	Bias	0.000	0.001	0.000	-0.000	-0.011	-0.009	0.000	-0.039	-0.007
		SE	0.004	0.005	0.005	0.004	0.011	0.012	0.004	0.007	0.008
120	20%	Bias	0.000	-0.000	-0.000	-0.000	-0.010	-0.010	0.000	-0.041	-0.009
		SE	0.003	0.003	0.004	0.003	0.007	0.007	0.004	0.007	0.007
30	33%	Bias	0.000	0.000	0.001	-0.000	-0.020	-0.016	-0.000	-0.070	-0.015
		SE	0.005	0.005	0.006	0.004	0.020	0.021	0.005	0.017	0.022
60	33%	Bias	0.000	0.000	0.001	-0.000	-0.020	-0.018	-0.000	-0.069	-0.015
		SE	0.003	0.003	0.004	0.003	0.014	0.015	0.004	0.012	0.014
120	33%	Bias	0.000	0.000	0.000	-0.000	-0.019	-0.018	-0.000	-0.069	-0.015
		SE	0.003	0.003	0.003	0.003	0.010	0.010	0.003	0.008	0.010
30	50%	Bias	0.000	0.000	0.001	-0.000	-0.029	-0.025	0.000	-0.096	-0.016
		SE	0.004	0.004	0.005	0.004	0.028	0.031	0.004	0.012	0.013
60	50%	Bias	0.000	0.000	0.001	-0.000	-0.029	-0.028	-0.000	-0.097	-0.016
		SE	0.003	0.003	0.003	0.003	0.016	0.016	0.003	0.008	0.009
120	50%	Bias	0.000	-0.000	0.000	0.000	-0.028	-0.026	0.000	-0.097	-0.016
		SE	0.002	0.002	0.002	0.002	0.012	0.011	0.002	0.006	0.006

Table 3
Standard Errors of Equating Under Parallel Forms and Across Levels of Anchor Representativeness.

Test Length	Anchor Percent	Anchor Representativeness: Representative			Non Representative (Moderate)			Non Representative (Severe)		
		Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt
30	20%	0.081	0.081	0.081	0.081	0.084	0.085	0.081	0.081	0.087
60	20%	0.057	0.057	0.057	0.057	0.062	0.064	0.057	0.057	0.060
120	20%	0.040	0.040	0.040	0.039	0.043	0.043	0.040	0.041	0.044
30	33%	0.079	0.079	0.080	0.079	0.093	0.097	0.079	0.085	0.098
60	33%	0.056	0.056	0.056	0.055	0.066	0.067	0.055	0.061	0.068
120	33%	0.039	0.039	0.039	0.039	0.046	0.047	0.039	0.044	0.048
30	50%	0.077	0.077	0.077	0.077	0.102	0.107	0.077	0.078	0.085
60	50%	0.054	0.054	0.054	0.054	0.069	0.070	0.054	0.058	0.060
120	50%	0.038	0.038	0.038	0.038	0.049	0.049	0.038	0.044	0.042

Table 4.
Means and Standard Deviations of Sample Equating RMSE Under Levels of Form Parallelism.

Test Length	Anchor Percent		Form Parallelism: Parallel — Unit Wt	Parallel — Prop Wt	Parallel — Info Wt	Nonparallel (Minor) — Unit Wt	Nonparallel (Minor) — Prop Wt	Nonparallel (Minor) — Info Wt	Nonparallel (Major) — Unit Wt	Nonparallel (Major) — Prop Wt	Nonparallel (Major) — Info Wt
30	20%	Mean RMSE	0.081	0.081	0.082	0.081	0.083	0.084	0.082	0.085	0.085
		SD RMSE	0.006	0.006	0.006	0.006	0.006	0.007	0.006	0.006	0.006
60	20%	Mean RMSE	0.057	0.057	0.057	0.058	0.059	0.059	0.059	0.064	0.061
		SD RMSE	0.003	0.003	0.004	0.004	0.004	0.004	0.003	0.004	0.004
120	20%	Mean RMSE	0.040	0.040	0.040	0.041	0.041	0.041	0.043	0.050	0.044
		SD RMSE	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.002
30	33%	Mean RMSE	0.079	0.079	0.080	0.080	0.082	0.082	0.081	0.090	0.082
		SD RMSE	0.005	0.005	0.005	0.006	0.006	0.006	0.005	0.006	0.005
60	33%	Mean RMSE	0.056	0.056	0.056	0.056	0.058	0.058	0.058	0.073	0.059
		SD RMSE	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.005	0.003
120	33%	Mean RMSE	0.039	0.039	0.039	0.040	0.041	0.041	0.043	0.060	0.045
		SD RMSE	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.003
30	50%	Mean RMSE	0.077	0.077	0.077	0.078	0.082	0.083	0.0[illegible]	0.102	0.082
		SD RMSE	0.005	0.005	0.005	0.005	0.005	0.006	0.005	0.008	0.006
60	50%	Mean RMSE	0.054	0.054	0.054	0.055	0.057	0.058	0.057	0.089	0.059
		SD RMSE	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.005	0.003
120	50%	Mean RMSE	0.038	0.038	0.038	0.039	0.041	0.041	0.042	0.082	0.044
		SD RMSE	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.002

Table 5
Means and Standard Deviations of Sample Bias in Equated Scores Under Levels of Form Parallelism.

Test Length	Anchor Percent		Form Parallelism: Parallel — Unit Wt	Parallel — Prop Wt	Parallel — Info Wt	Nonparallel (Minor) — Unit Wt	Nonparallel (Minor) — Prop Wt	Nonparallel (Minor) — Info Wt	Nonparallel (Major) — Unit Wt	Nonparallel (Major) — Prop Wt	Nonparallel (Major) — Info Wt
30	20%	Bias	0.000	0.002	0.004	0.000	0.002	0.004	-0.000	-0.026	0.008
		SE	0.006	0.006	0.007	0.005	0.009	0.011	0.005	0.009	0.011
60	20%	Bias	-0.000	-0.001	0.000	0.000	0.004	0.005	-0.000	-0.025	0.007
		SE	0.004	0.005	0.005	0.005	0.006	0.007	0.004	0.006	0.006
120	20%	Bias	0.000	-0.000	-0.000	-0.000	0.004	0.005	0.000	-0.025	0.007
		SE	0.003	0.003	0.004	0.003	0.005	0.005	0.003	0.005	0.005
30	33%	Bias	0.000	0.000	0.001	-0.000	0.005	0.006	-0.000	-0.046	0.009
		SE	0.005	0.005	0.006	0.004	0.009	0.010	0.004	0.008	0.009
60	33%	Bias	-0.000	-0.000	0.001	0.000	0.005	0.006	-0.000	-0.045	0.008
		SE	0.003	0.003	0.004	0.004	0.007	0.007	0.004	0.006	0.006
120	33%	Bias	0.000	0.000	0.000	0.000	0.005	0.007	0.000	-0.041	0.012
		SE	0.003	0.003	0.003	0.003	0.005	0.005	0.003	0.005	0.006
30	50%	Bias	0.000	0.000	0.001	0.000	0.010	0.012	0.000	-0.068	0.013
		SE	0.004	0.004	0.005	0.003	0.012	0.013	0.004	0.010	0.010
60	50%	Bias	0.000	0.000	0.001	0.000	0.009	0.010	0.000	-0.068	0.013
		SE	0.003	0.003	0.003	0.003	0.007	0.008	0.003	0.006	0.007
120	50%	Bias	-0.000	0.000	0.000	0.000	0.010	0.010	0.000	-0.068	0.012
		SE	0.002	0.002	0.002	0.002	0.005	0.005	0.002	0.005	0.005

Table 6
Standard Errors of Equating Under Levels of Form Parallelism

Test Length	Anchor Percent	Form Parallelism: Parallel			Nonparallel (Minor)			Nonparallel (Major)		
		Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt
30	20%	0.081	0.081	0.081	0.081	0.083	0.084	0.082	0.081	0.084
60	20%	0.057	0.057	0.057	0.058	0.058	0.059	0.059	0.058	0.060
120	20%	0.040	0.040	0.040	0.041	0.041	0.041	0.043	0.043	0.044
30	33%	0.079	0.079	0.080	0.080	0.081	0.082	0.081	0.078	0.082
60	33%	0.056	0.056	0.056	0.056	0.058	0.058	0.058	0.057	0.059
120	33%	0.039	0.039	0.039	0.040	0.041	0.041	0.043	0.043	0.044
30	50%	0.077	0.077	0.077	0.078	0.082	0.082	0.079	0.076	0.081
60	50%	0.054	0.054	0.054	0.055	0.057	0.057	0.057	0.057	0.058
120	50%	0.038	0.038	0.038	0.039	0.040	0.040	0.042	0.045	0.042

Table 7
Correlations of Anchor Score and Equated Score With Expected Base Form Score Under Parallel Forms and Across Levels of Anchor Representativeness.

Test Length	Anchor Percent	Anchor Representativeness: Representative			Non-Representative (Moderate)			Non-Representative (Severe)		
		Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt
30	20%	0.902	0.901	0.900	0.901	0.896	0.893	0.902	0.894	0.891
60	20%	0.948	0.948	0.947	0.948	0.940	0.937	0.948	0.944	0.943
120	20%	0.974	0.974	0.974	0.974	0.971	0.970	0.974	0.971	0.969
30	33%	0.906	0.906	0.905	0.906	0.879	0.870	0.906	0.882	0.870
60	33%	0.950	0.950	0.950	0.951	0.935	0.932	0.951	0.936	0.931
120	33%	0.975	0.975	0.975	0.975	0.966	0.965	0.975	0.967	0.964
30	50%	0.911	0.911	0.911	0.912	0.866	0.854	0.912	0.896	0.896
60	50%	0.953	0.953	0.953	0.954	0.929	0.927	0.953	0.943	0.944
120	50%	0.976	0.976	0.976	0.976	0.963	0.962	0.976	0.971	0.972

Table 8
Correlations of Anchor Score and Equated Score With Expected Base Form Score Under Levels of Form Parallelism.

Test Length	Anchor Percent	Form Parallelism: Parallel			Nonparallel (Minor)			Nonparallel (Major)		
		Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt	Unit Wt	Prop Wt	Info Wt
30	20%	0.902	0.901	0.900	0.900	0.897	0.895	0.898	0.894	0.895
60	20%	0.948	0.948	0.947	0.947	0.946	0.945	0.943	0.941	0.942
120	20%	0.974	0.974	0.974	0.973	0.972	0.972	0.969	0.968	0.969
30	33%	0.906	0.906	0.905	0.905	0.902	0.900	0.902	0.897	0.900
60	33%	0.950	0.950	0.950	0.949	0.947	0.947	0.946	0.943	0.945
120	33%	0.975	0.975	0.975	0.973	0.973	0.973	0.970	0.969	0.969
30	50%	0.911	0.911	0.911	0.909	0.903	0.901	0.907	0.899	0.903
60	50%	0.953	0.953	0.953	0.951	0.949	0.949	0.949	0.945	0.947
120	50%	0.976	0.976	0.976	0.974	0.974	0.974	0.971	0.970	0.971

Figure 1 Means of Sample Bias in Equated Scores Under Parallel Forms and Across Levels of Anchor Representativeness

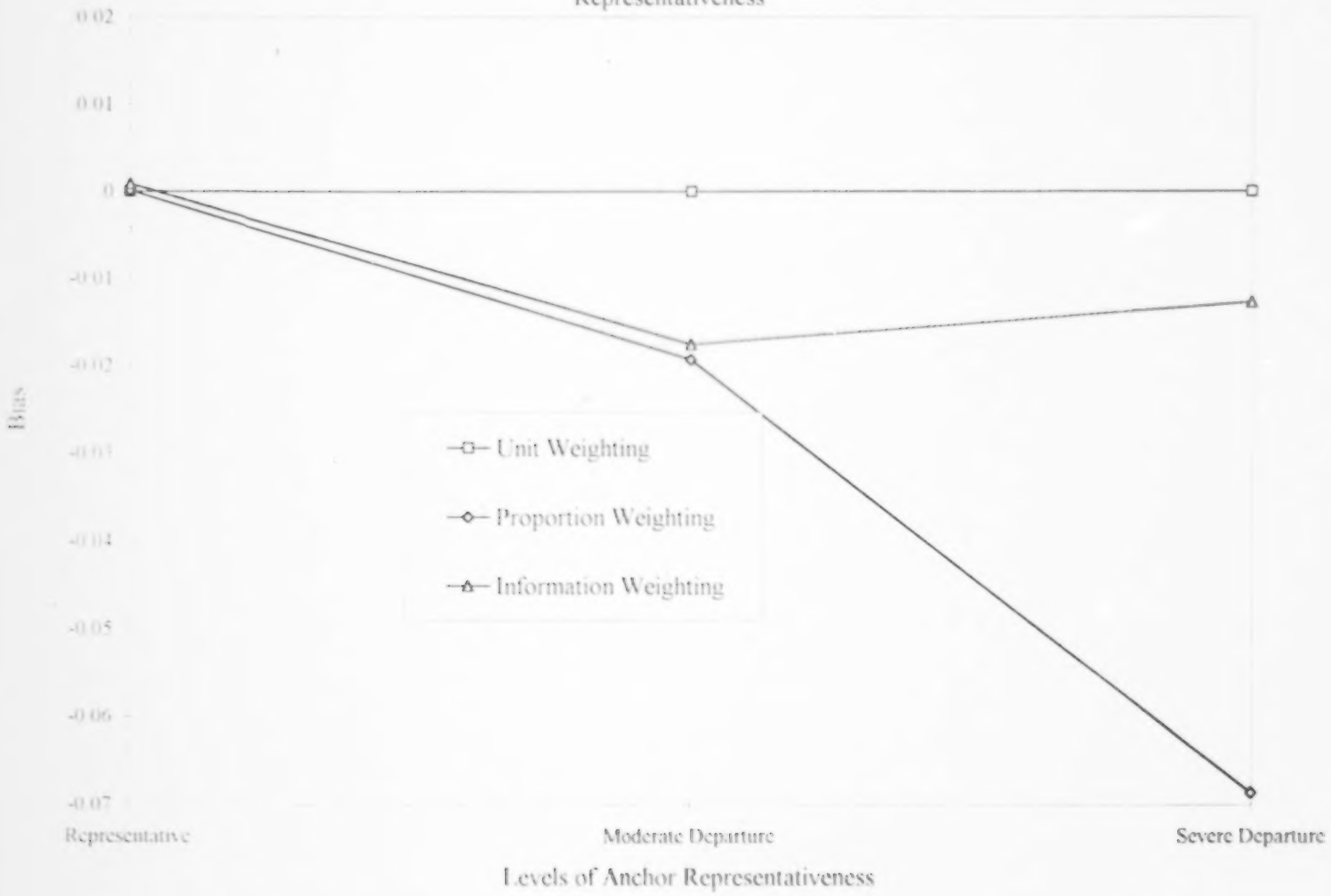

Figure 2 Standard Errors of Equating Under Parallel Forms and Across Levels of Anchor Representativeness

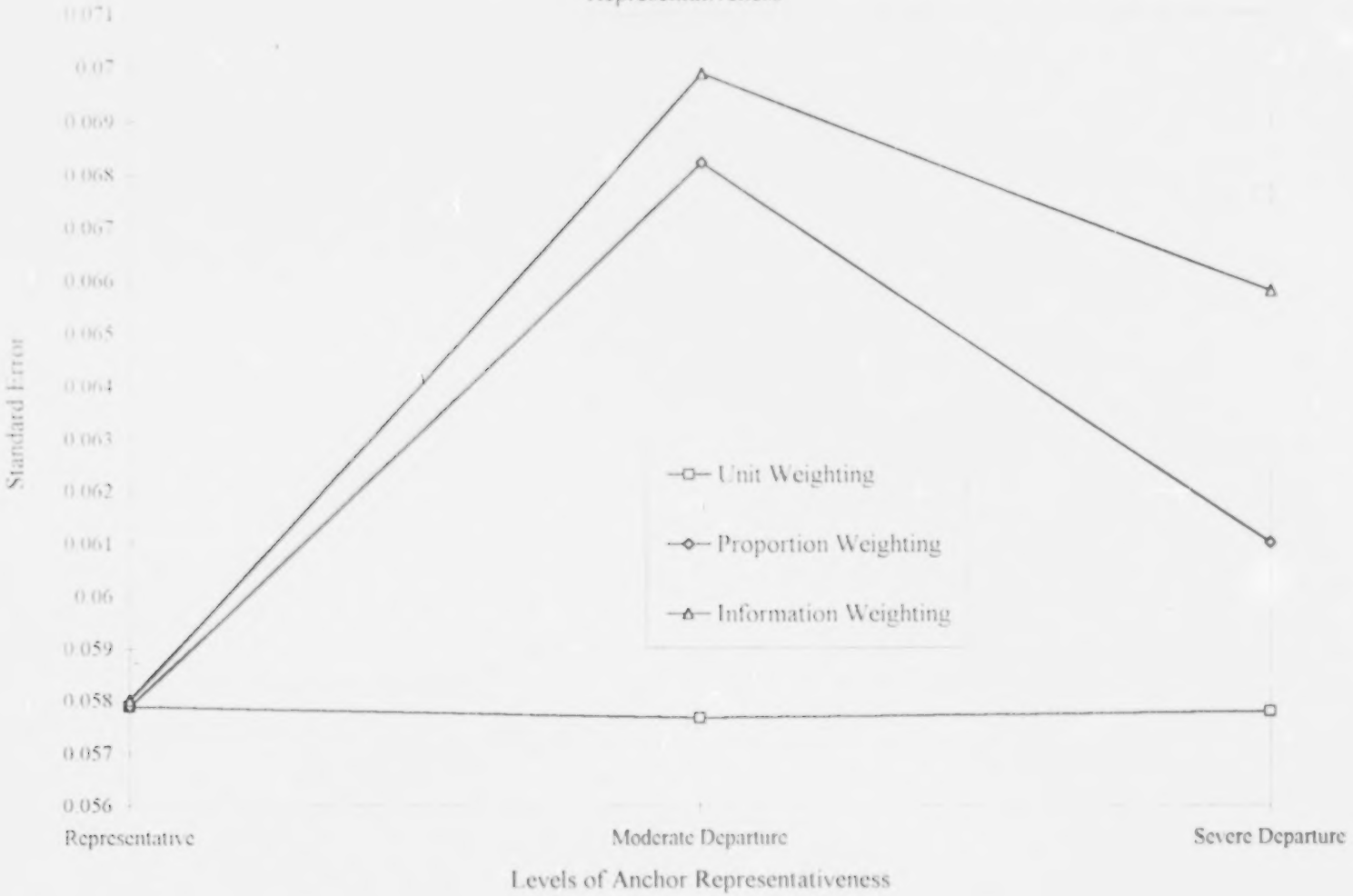

Figure 3 Means of Sample Bias in Equated Scores Under Parallel Forms and Across Levels of Form Parallelism

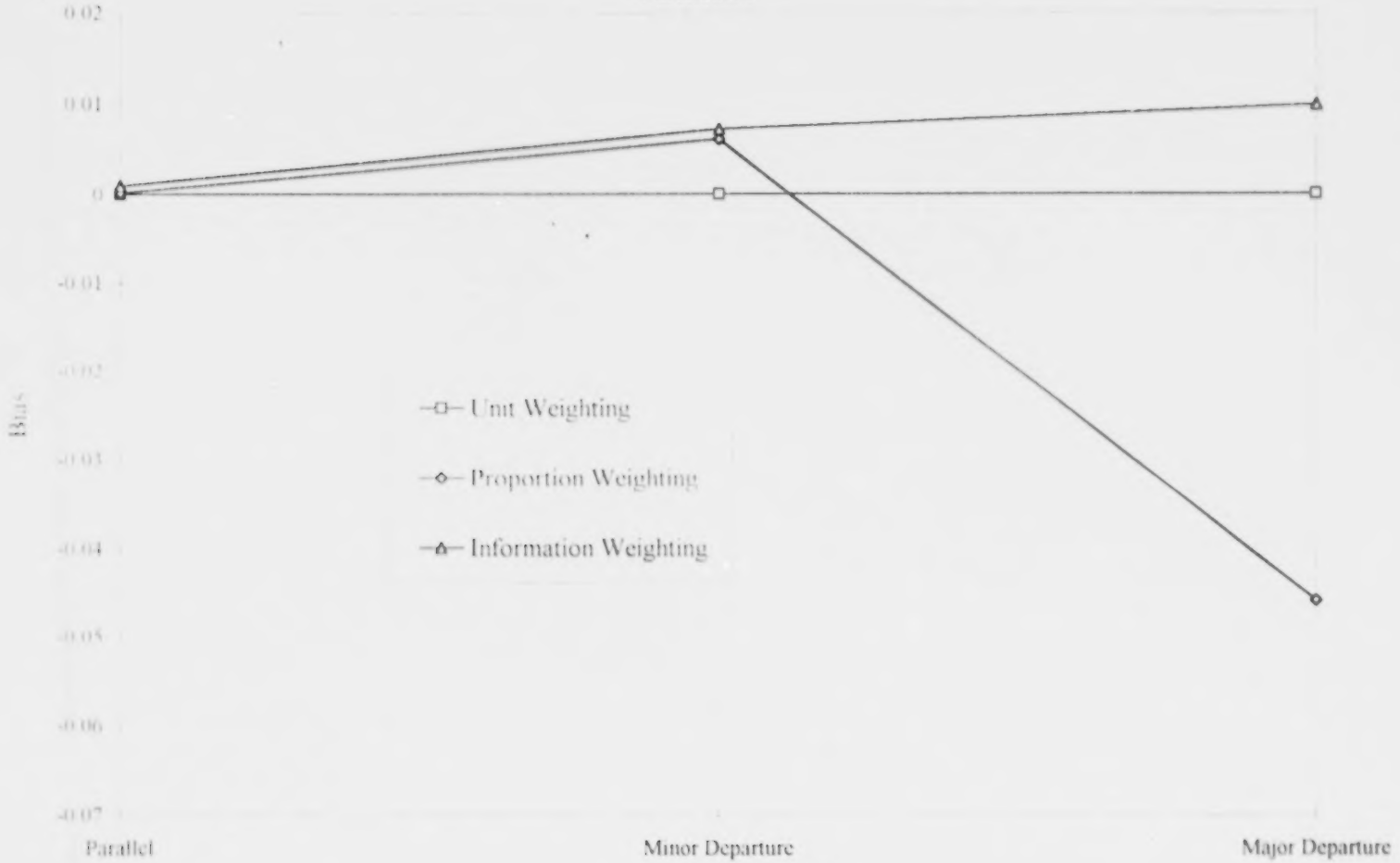

Figure 4 Standard Errors of Equating Under Levels of Form Parallelism